The Annals of Statistics
2007, Vol. 35, No. 6, 2358–2364
DOI: 10.1214/009053607000000433
Main article DOI: 10.1214/009053606000001523
© Institute of Mathematical Statistics, 2007

DISCUSSION: THE DANTZIG SELECTOR: STATISTICAL ESTIMATION WHEN *p* IS MUCH LARGER THAN *n*

BY BRADLEY EFRON[1], TREVOR HASTIE[2] AND ROBERT TIBSHIRANI[3]

Stanford University

1. Introduction. This is a fascinating paper on an important topic: the choice of predictor variables in large-scale linear models. A previous paper in these pages attacked the same problem using the "LARS" algorithm (Efron, Hastie, Johnstone and Tibshirani [3]); actually three algorithms including the Lasso as middle case. There are tantalizing similarities between the Dantzig Selector (DS) and the LARS methods, but they are not the same and produce somewhat different models. We explore this relationship with the Lasso and LARS here.

2. Dantzig selector and the Lasso. The definition of the Dantzig selector (DS) in (1.7) can be re-expressed as

$$\min_{\beta} \|X^T(y - X\beta)\|_{\ell_\infty} \quad \text{subject to} \quad \|\beta\|_{\ell_1} \le s. \tag{1}$$

This makes it look very similar to the Lasso (Tibshirani [6]), or basis pursuit (Chen, Donoho and Saunders [1]):

$$\min_{\beta} \|y - X\beta\|_{\ell_2} \quad \text{subject to} \quad \|\beta\|_{\ell_1} \le s. \tag{2}$$

With a bound on the ℓ_1 norm of β, Lasso minimizes the squared error while DS minimizes the maximum component of the gradient of the squared error function. If s is large so that the constraint has no effect, then these are the same. However, for other values of s, they are a little different; see Figure 1.

The least angle regression (LARS) algorithm (Efron, Hastie, Johnstone and Tibshirani [3]) for solving the Lasso path makes them look tantalizingly close (see also the homotopy algorithm of Osborne, Presnell and Turlach [4]). In LARS, we start with $\beta = 0$ and identify the predictor having maximal absolute inner product with y. We then increase/decrease its coefficient (depending on the sign of the inner product), which in turn reduces its absolute inner product with the current residual $r = y - X\hat{\beta}$. We continue until some other predictor has as large an absolute inner product with the current residual. That predictor is then included in the

Received January 2007.
[1]Suported in part by NSF Grant DMS-00-72360 and by NIH Grant 8R01 EB002784.
[2]Suported in part by NSF Grant DMS-05-05676 and by NIH Grant 2R01 CA 72028-07.
[3]Suported in part by NSF Grant DMS-99-71405 and by NIH Contract N01-HV-28183.